On the Origin of Metadata

Erik Mannens, Ruben Verborgh, Seth van Hooland, Laurence Hauttekeete, Tom Evens, Sam Coppens, and Rik Van de Walle

In: Information, 3 (4), 2012.

http://www.mdpi.com/2078-2489/3/4/790/pdf

Information **2012**, *3*, 790-808; doi:10.3390/info3040790

OPEN ACCESS

information

ISSN 2078-2489

www.mdpi.com/journal/information

Article

On the Origin of Metadata

Erik Mannens [1,]*, Ruben Verborgh [1], Seth van Hooland [2], Laurence Hauttekeete [3], Tom Evens [3], Sam Coppens [1] and Rik van de Walle [1]

[1] ELIS-Multimedia Lab, Ghent University-iMinds, Ghent 9000, Belgium; E-Mails: ruben.verborgh@ugent.be (R.V.); sam.coppens@ugent.be (S.C.); rik.vandewalle@ugent.be (R.V.D.W.)

[2] Information & Communication Department, Université Libre de Bruxelles, Brussels 1050, Belgium; E-Mail: svhoolan@ulb.ac.be

[3] MICT, Ghent University-iMinds, Ghent 9000, Belgium; E-Mails: laurence.hauttekeete@ugent.be (L.H.); tom.evens@ugent.be (T.E.)

* Author to whom correspondence should be addressed; E-Mail: erik.mannens@ugent.be; Tel./Fax: +32-9-331-4993.

Received: 3 August 2012; in revised form: 3 December 2012 / Accepted: 4 December 2012 / Published: 7 December 2012

Abstract: Metadata has been around and has evolved for centuries, albeit not recognized as such. Medieval manuscripts typically had illuminations at the start of each chapter, being both a kind of signature for the author writing the script and a pictorial chapter anchor for the illiterates at the time. Nowadays, there is so much fragmented information on the Internet that users sometimes fail to distinguish the real facts from some bended truth, let alone being able to interconnect different facts. Here, the metadata can both act as noise-reductors for detailed recommendations to the end-users, as it can be the catalyst to interconnect related information. Over time, metadata thus not only has had different modes of information, but furthermore, metadata's relation of information to meaning, *i.e.*, "semantics", evolved. Darwin's evolutionary propositions, from "species have an unlimited reproductive capacity", over "natural selection", to "the cooperation of mutations leads to adaptation to the environment" show remarkable parallels to both metadata's different modes of information and to its relation of information to meaning over time. In this paper, we will show that the evolution of the use of (meta)data can be mapped to Darwin's nine evolutionary propositions. As mankind and its behavior are products of an evolutionary process, the evolutionary process of metadata with its different modes of information is on the verge of a new-semantic-era.

Keywords: metadata; evolution; information; semantics; Darwin

1. Introduction

In recent years, the World Wide Web has become the victim of its own success. According to Google, every two days now we produce as much information as we did from the dawn of civilization up to 2003. One of the problems is that the Internet was initially designed as an instrument to share documents, and not intended as a platform for sharing the information that is enclosed within these documents. Tim Berners-Lee himself was one of the first to suggest the idea of unleashing the real meaning of all the data out there [1]. However, if one wants to build computers that are able to interpret such vast information streams, one has to explain in detail the relation of information to meaning, *i.e.*, "semantics", of the data in the information streams to these machines. In neo-Darwinian theory, the only relevant information in DNA was thought to be that which codes for proteins/genes. The rest was thought to be "junk" leftovers from the past, or useless duplicates of currently functional genes. However, when we look beyond the Shannon definition of information, we come across an entirely different kind of information that is called *metadata*. In this paper we will show that the evolution of the use of (meta)data can be mapped remarkably well to Darwin's Theories [2–4]. As mankind and its behavior are products of an evolutionary process, the evolutionary process of metadata with its different modes of information is on the verge of a new-semantic-era. Only when we acknowledge the relevance of the Darwinian metadata evolution, will we be able to grasp the vast information streams out there, give it meaning and safeguard this for future generations. "If content is King, then metadata is Queen", *i.e.*, when data volumes stack up, the importance of metadata—according to the ignorant the "junk" leftovers—and its semantics will only improve. Indeed, links between data—*i.e.*, metadata—become more important than the data itself. Even more data is a good thing, as long as the bulk of this "more data" is metadata.

2. Proposition I: Species Have an Unlimited Reproductive Capacity

2.1. Darwin's View

Darwin's first argument starts from one simple observation, *i.e.*, every living species theoretically has an unlimited reproductive capacity. He calculated at the time that one pair of elephants—being known as one of the least fertile animals around—would create an offspring of about 15 million within five centuries [2]. His point being that if there would be no hindering of the reproductive capacity at all, a gigantic number of descendants will emerge in rather a short time span. Try answering the following question intuitively: If mankind would continue to grow as it does now at a rate of about 80 million people a year, how long would it take to fill up the entire universe (diameter ± 18 billion light years) with human flesh? A mere 4830 years [5]! We humans tend to vastly underestimate the accretion of exponential arithmetic progressions.

2.2. The (Meta)Data Counterpart

Let's consider from here on the (meta)data counterpart of a *species* as a collection of data—and thus information—of a single type being able to be exchanged between humans and/or computers. In the early days of non-verbal lasting communication, let's say +40,000 years ago, pre-historic humans had only murals at their disposal to depict information. Needless to say, the exchange of this information targeted a small in-crowd, *i.e.*, the few Neanderthals that visited each others caves. In addition, this time consuming manual labor exerted to share information lasted well into the middle ages, when an average monk could copy about one book per year. Mind you, at that time, data and metadata were already intertwined, albeit not recognized as such. Medieval manuscripts typically had illuminations at the start of each chapter, being both a kind of signature for the author writing the script as also being a pictorial chapter anchor for the illiterates at the time. The invention of the printing press in the 15th century was a first major milestone for data replication for the masses. Since then, a first huge wave of "hard copy" data and information, e.g., through books and daily newspapers, flooded our world. Needless to say, librarians have tried to be on top of that huge pile of information *via* all sorts of—mostly manual—indexing schemes the last couple of hundred years or so. Then, at the end of the 20th century, the digital era in the form of the Internet dawned to spawn an even bigger wave of data overflow. Back in 2008 Google stopped counting unique URLs at 1 trillion (1,000,000,000,000) and every digital native is now producing over 10K digital artifacts per year. The Internet of Things is perhaps still a decade away, but imagine every digital appliance out there—be it a light bulb, refrigerator, or mobile phone—being able to constantly gather and publish data online. As stated earlier, we humans tend to vastly underestimate the accretion of exponential arithmetic progressions. As such, the solution to information overload is more information, as long as this information will be metadata [6].

3. Proposition II: In Fact, Numbers of Each Species are Limited

3.1. Darwin's View

As previously stated, *purely theoretically*, all species can breed themselves amazingly fast to vast numbers, but *in fact*, the numbers of each species stay relatively stable over long periods of time. Moreover, definitive extinction of a species is the normal course of events, *i.e.*, 99.9% of all species that ever lived are extinct [7]. Since the Cambrian explosion, our planet repeatedly went through periods of high mortality, e.g., due to the impact of large asteroids that caused abrupt climate changes. In his second proposition, Darwin called these counterforces of unbridled expansion of species the "checks on growth" [2]. From observations in his own garden, he could explain the decrease (increase) in bird populations over a decade due to (the lack of) the following natural "checks on growth": food shortages, natural disasters, predators and/or epidemics.

3.2. The (Meta)Data Counterpart

EU countries produce over five billion electronic documents per year—which is huge, but a lot of it will eventually be lost. Within this digital age, we also have "checks on growth". Natural disasters (like flooding and fires) still happen, but it's the aging and non-interoperability of IT-infrastructure, (like bitrot and old proprietary formats) and the lack of adequate back up solutions that are the biggest threads to insurmountable data loss now. Therefore, preservation of digital objects must be done on three conceptual metadata levels [8]. One must preserve the *medium*, the *technology*, and the *intellectual content*.

On the lowest level, (***preserving the medium***) a digital file consists of bits and bytes saved on hardware systems, which are liable to wear-and-tear. These hard disks and tapes have indeed a limited life span as their bit streams can be altered by external influences, e.g., corruption of these digital carriers. This level therefore needs error correcting hardware and software solutions. The authenticity of digital data is harder to guarantee and maintain than it is for analogue data. In the latter case, it is sufficient to describe all features of the physical object, but for digital information the whole provenance and continuous processing must be archived too.

On a higher level (***preserving the technology***), file formats and compression formats, e.g., PDF (for text), MPEG-4 (for video), MP3 (for audio), and JPEG (for images), describe the way the bits can be transformed to an interpretable multimedia representation. When a file format becomes obsolete, one has two options to preserve the stored data: *migration*, *i.e.*, moving electronic files from one application to another, or *emulation*, *i.e.*, designing software and/or hardware that will mimic a specific application. Both have pros and cons, as migration can cause data loss and emulation can become very complex. In either case, such metadata is needed to support these actions. At the same time, open standards are vital to foster future understandability and migratebility of file formats, as the interpretation of proprietary file formats needs proprietary software, which is again more difficult to keep track of in a sustainable manner. At this level, it is also very important to preserve the look and feel of multimedia objects, as, e.g., resolution or color values might change when migrating file formats. Thus, a rich description of the look and feel is also necessary.

On the highest level (***preserving the intellectual content***), the information should always remain interpretable. Institution structures, terminologies, the designated community, and the rights of an object or institution might change over time. To keep that kind of information interpretable too, enough extra "semantic" information should be included in the information package. At this level, the digital *species* not only needs descriptive metadata for a general description of the object, but also rights metadata and contextual metadata for describing the relations of the content information to information which is not packed in the information package itself. Examples of such contextual metadata are related datasets, references to documents in the original environment at the moment of publication, and helper files. This contextual information becomes indispensable overtime, as the initial providers of the extra information on the datasets themselves are no longer available to explain why they archived a dataset in a certain way. As such, a dataset should provide enough contextual metadata to keep it interpretable for a designated community without the help of external experts.

As such, if one fails on any of those three levels, data will be lost and thus its accompanying information will be extinct. No descriptive metadata means no web crawler will pick it up, hence no search engine can return it, and hence it does "not exist" for most of the entire world population, *i.e.*, it is "virtually" intellectual extinct. Overtime, it can now only be read by a small in-crowd. No technical metadata means no software will be able to interpret it within two decades, and hence it becomes merely "medium" not able to be interpreted anymore, *i.e.*, it is now also technologically extinct. From then on, it is merely a question of time that also the medium itself degenerates *via*, *i.e.*, bitrot, to make the medium itself extinct too.

4. Proposition III: Individual Variation & Differences are Hereditary

4.1. Darwin's View

Darwin sensed that the chances of survival—and thus reproduction—of an organism are determined for the most part by the intrinsic characteristics of the individual itself, more than they are by pure coincidence. He also noted—as he carefully observed his own dovecot for years on end—that individuals of the same species differ from each other and that these differences are transferred from parent to child. He called these differences "natural variations". To Darwin, however, the existence of hereditary characteristics was an essential assumption. If inheritance did not exist, the evolutionary mechanism that he believed to have discovered could never have worked. In 1866, not long after Darwin's *Origin* was published, an article written by Gregor Johann Mendel—a Bohemian Augustinian monk—appeared in an obscure journal in Brünn [9]. Darwin died in 1882 without ever having heard of Mendel. However, this article contained the basic laws of hereditary transmission and the idea that something like "genes" had to exist, so hereditary characteristics—unchanged and as separate entities—are transferred from generation to generation.

4.2. The (Meta)Data Counterpart

Just looking at a small sub-set of all the multimedia data out there, let's say the images, there are already a myriad ways of describing extra information—from low level to high level—for this *species*, *i.e.*, with the standard EXIF, MPEG7, XMP, i3a, IPTC, *etc.* We could call them all "natural variations". According to the Open Archival Information System [8] all *natural variations* need different types of metadata to fully describe the information of their assets for lasting reproducing capacity. As stated above, data can be lost on all three conceptual levels (medium, technology, and content). It is through (painful) evolution that we now pinpointed all necessary kinds of metadata out there, so we are able to make sense of the "real" data and are sure to safeguard this information and knowledge for future generations to come. As such, the following six types of metadata—from now on the *hereditary characteristics*—are a guarantee that the data can be sufficiently described to fully satisfy these three levels:

> ***Binary metadata*** describe the data on bit level. Bitstreams are the actual data in a file. Binary metadata, e.g., file system information and file header information, keep the enclosed information accessible by pointing out how the bits should be transformed to a representation of the data, e.g., in a certain compression format.

> ***Technical metadata*** describe the data on file level. Data formats and their derivatives evolve quickly. As both container and compression formats age, it is hard to find software that is still able to interpret old formats. The only way to keep this kind of information accessible is to support migration and/or emulation in which the technical metadata, e.g., coder-decoder (codec) information, will be key in keeping that possible.

Structural metadata describe the relationships between a set of files that correspond to a possible representation of the intellectual content of certain data. A certain book might be an aggregation of a set of chapters with pages in a specific order identified by the table of contents. This structural metadata is necessary to fully describe the complete book as a correct ordering of these pages.

Descriptive metadata describe extra data, e.g., author, title, location, date, *etc.*, to better find and locate the original data. When exchanging digital multimedia content from different industries/institutions—be it broadcasters, libraries, cultural institutions, and archives—an additional problem concerning descriptive metadata arises, *i.e.*, a lot of industries/institutions already describe, control, and save their descriptive metadata according to their own (standardized) schemes. As such, some file these extra metadata as metadata, others file them as real data. Both strategies have their pros and cons. If a coordinating institution wants to file these extra descriptions as metadata, it means it is forced to choose one metadata standard to do so, which is not obvious, as most metadata schemes are domain specific. To guarantee lossless filing of all descriptive metadata, our coordinating institution must opt for the lowest common denominator of all descriptive metadata schemes used by all partnering industries/institutions, which would lead to an enormous unmaintainable metadata scheme. It is therefore best to archive the descriptive metadata (in its original metadata format) together with their original data, thus being sure not to lose any information ever.

Preservation metadata describe essential extra data that support and document the digital preservation process. No digital storage device is perfect and perpetually liable, as bit preservation is still an unsolved paradigm. The simplest model of these failures is analogous to the decay of radioactive atoms. Each bit in a data file independently is subject to a random process that has a constant small probability per unit time of causing its value to flip. The time after which there is a 50% probability that a bit will have flipped is the *bit half-life*. The requirement of a 50% chance that 1 petabyte of data will survive for a century translates into a required bit half-life of 8×10^{17} years. To put things into perspective, the current estimate of the age of the universe U is 1.4×10^{10} years, so this is a bit half-life of approximately only 6×10^{7} U [10]. As stated earlier, information in a digital form is a conceptual object. This information can be altered and copied pretty easily without one notifying that in its visible representation. Opposed to analogue information, it is indeed much harder to preserve the authenticity of digital information. This too can be solved by adding tenability metadata to the preservation package of the archived essence. Such metadata have check sums, digital signatures, certificates, encryption, and cyclic redundancy checksum for indicating the data is not altered without it being documented. Furthermore, an archived dataset also needs its provenance documented. This type of preservation metadata (e.g., encoding software, version history, references to the original sources, *etc.*) describes the genesis of the intrinsic information, *i.e.*, the original owners of the data, the processes determining the current form of that data, and all of its available, intermediate versions, as this information is vital in verifying all changes the data has experienced from genesis until date. Lastly, context-aware metadata (e.g., related data sets, help files, original language on first publication, *etc.*) must be retained, as these describe possible relationships of the intrinsic data with other data that is not embraced within its own information package.

Lastly, ***rights metadata*** describe the rights on digital objects (e.g., rights metadata for describing copyright statements, (changing) licenses, and possible grants that are given), as this info is also vital to guarantee long-term access to the data, and thus must be saved too.

5. Proposition IV: Survival of the Fittest

5.1. Darwin's View

Some genetic differences do not affect the survival and reproduction chances of the individual, but others do largely determine these chances. This means that the counter-forces that restrain the numbers of each species, work selectively. They are—as it were—"elected" by nature, thus these "fit" characteristics (both purely physical, and behavioral) are cumulative being transferred generation after generation. Darwin called this selection of characteristics "natural selection" and he clarified that mere coincidence is thus eliminated over long periods of time. Darwin considered it obvious that a "fast" doe had a much better chance of survival—and thus reproduction—than a "slow" doe, which would be foiled much earlier by some predator. Darwin further noted that internal competition is also a driving force in evolution. Within this context, biologists call it the "Red Queen" [11] effect. Conspecifics are competitors and any improvement in the genome of one individual forces the other members of that species also some kind of amelioration. Sometimes evolution is driven by some kind of race between species. Plants have developed all kinds of poison to prevent them from being eaten by herbivores. For their part, herbivores developed enzyme systems that can degrade these plant toxins. In other words, there is a constant race in progress between members of the same species and between members of different species. He, who dares to stand still, will—compared to his opponents—automatically go backwards.

5.2. The (Meta)Data Counterpart

If we have built two competing solutions to a data problem both *genetically* different with one being a closed proprietary solution and the other being an open standardized solution, which of the two is likely to be *(s)elected* by nature and thus has the *fittest* characteristics? The answer to that question is mostly dependent on that other driving factor, *i.e.*, internal competition.

The bigger the market share of a certain software solution, the bigger the change as being the "de facto" standard, and the bigger the change in moving forward using its own proprietary solutions. Reactive, agile contenders tend to develop open solutions, using open standards which are platform independent, universally usable and vendor neutral. These open standards are subject to full public assessment and use without constraints in a manner equally available to all parties, thus increasing the speed and uptake by a bigger crowd. As such, open standards have proven to be an important facilitator for innovation. By providing an agreed, reliable and globally valid base of technology, open standards allow innovators to develop highly competitive, innovative technologies and solutions "on top" of that standard. At the same time, they have got some safeguards regarding the potential for global market outreach. The most prominent example from the last decade is the World Wide Web itself. Having open standards, publicly available on royalty-free terms, was the base for a wide wave of innovation which has, in fact, revolutionized the way in which we live, operate, and communicate. Open standards have boosted innovation and growth.

When such agile open solutions get enough momentum, *i.e.*, get a big enough percentage of the market share, three things can happen:

1. The big, proprietary player buys the agile open solution and incorporates it within its own software solution. The big player embraces the obviously *fit* characteristics of the open agile solution and continues its world domination, e.g., Google bought Freebase and now improves its search results using a semantic knowledge graph [12].
2. The big, proprietary player buys the agile open solution, puts it aside and hopes to maintain its world domination without embracing evolution, e.g., SUN Microsystems bought Netscape's web server suite—at that time superior to SUN's suite, but failed to incorporate it within its own suite and finally stopped further developing evolutions of it as it tried to win the Java battle [13]. In two years time Microsoft's IIS server took over, together with an incumbent agile open source software solution, the Apache software foundation [14]. He, who dares to stand still, will—compared to his opponents—automatically go backwards and be eradicated.
3. The small goes for world domination, becomes big itself and than the use of "open standards" becomes a subtle game changer in just getting on top of one another. Before 2005, when Apple was far smaller than Microsoft, it invested heavily in "open standards" through W3C, *i.e.*, HTML, CSS, and JavaScript, and they told at numerous occasions that Microsoft was holding back evolution through not adopting these standards as proposed, but keep on building their own proprietary version of it [15]. Only five years later, the situation completely reversed. Now Apple is more dominant than Microsoft, pushing its own end-to-end ecosystem without adhering to all "open standards" out there. Microsoft took up standardization again and had one of the first browsers implementing most of the new HTML5 features.

As such, open standards are to be key to further protrude evolution. They instigate a future-proof paradigm to fulfill the necessary conditions for simple, reliable communication between systems, processes, and humans. As such, choosing open standards is highly strategic. Their benefits and positive impact are debated and seen at the highest decision making levels. Interoperability is a major requirement for data and knowledge exchange as societies, governments and industry increasingly move towards global collaboration and integration. Open standards built on the principles of openness, transparency, and consensus lay the grounds for innovation and growth, for flexibility and choice, for global market success, and fair competition. In other words, open standards is where society, government, and industry align and where—in the end—everyone will benefit, hence sure to be *(s)elected* by nature and thus *hereditary* ever after.

Relating this all back to just metadata standards: Since 2008 W3C's Media Annotation working group developed a schema to facilitate cross-community data integration of information related to media objects in the Web—such as video, audio and images. In 2011, Microsoft, Yahoo, and Google came up with an alternative solution, *i.e.*, schema.org. Since then W3C not only tries to incorporate schema.org within their Media Annotations group, but also made some additions and improvements to schema.org [16]. As we will see later on, standards can be developed on different levels of granularity, *i.e.*, ranging from mere definition of syntax, over description of a data model, definition of a vocabulary, to the attaining of semantics.

6. Proposition V: Environment is a Selecting Mechanism

6.1. Darwin's View

During his lifetime, however, Darwin was not able to come up with his own concrete examples of his "Survival of the Fittest" theory. In the second half of the 19th century, a few butterfly collectors were able to do so without them even knowing it. In 1848, a rare variant of the Biston Betularia—a light grey moth—was discovered in Manchester, *i.e.*, the Biston Carbonaria, raven black in colour. The following decades this Biston Carbonaria was more and more noticed, in so far that around 1900, the Biston Betularia became a rarely spotted specimen. Good to know that the habitat of both moth species is indeed the silver birch. Where the Biston Carbonaria was always eaten pre 1850, about 1900 it was invariably the Biston Betularia that was eaten because the bark of the silver birch was covered with a thick layer of soot due to the industrial revolution at that time. Nowadays it is again the Biston Carbonaria that is about to vanquish [5]. Darwin already understood very well that there is no absolute definition of a "fit" gene. What fits, is determined by the environment. If the environment changes, the definition of what "fit" genes are, is then automatically changed as well.

6.2. The (Meta)Data Counterpart

The last half-millennium information experts have tried to be on top of the huge pile of "hard copy" information via all sorts of—mostly manual—card indexing schemes. This worked remarkably well into the first half of the 20th century. However, with the advent of both advanced electronic printing equipment and the digital age, "hard copy" and especially "soft copy" information exponentially grew, that it could even not be indexed anymore via card indices. Suddenly, its characteristics were not "fit" anymore. Its only fit characteristic, *syntax*, was just not enough to scale in the current Internet information space. It is not easy adjustable, nor is it easily interchangeable, let alone have embedded semantics, *i.e.*, all characteristics needed to *survive* in the current information space. Mind you, it is not always the current *best* (meta)data standard that survives. As said, the environment itself determines what fits best. Remember the video recording format standards battle in the late seventies of the last century between JVC and Sony? Albeit Sony hit the consumer market first and had the Betamax standard that was definitely superior in quality, it was JVC's VHS standard that prevailed, as Sony had a very lousy marketing campaign and above all too strict licensing schemes [17]. As such, Sony drove the major movie industries to the inferior VHS of JVC, which seemed "fit" enough for the purposes they had in mind at that time.

7. Proposition VI: Sexual Selection

7.1. Darwin's View

In 1871 Darwin published an important addition to his original version of the theory of evolution: *Selection in Relation to Sex* [3]. Apart from "blind" natural selection, evolution is also driven by "picky" sexual selection. Why on earth do peacock hens invariably choose the male peacock with the biggest and most beautiful tail, while the gaudy tail makes the bird less maneuverable and its garish colors easily attract predators? Almost all species that sexually reproduce themselves consist of two distinct sexes. And almost always members of the female sex produce few large gametes, whereas members of the opposite male sex small produce huge quantities of small gametes, *i.e.*, the largest cell in the human body is the female egg—production ready in 28 days—and the smallest actual cell is the male sperm cell—about 85k times smaller than an ovum, of which there are about 200 to 300 million present in each ejaculate. In pure physiological terms, the reproductive ability of a woman is thus very limited opposed to that of a man, *i.e.*, she has about six natural chances taking into account the limited time she is fertile and the time she has to invest of both being pregnant and taking care of the baby the first couple of years. As such women have a very good reason to cautiously deliberate *when* to deal with her limited reproductive opportunities and to carefully invest time *who* to select as mating partner. Evolutionary logic predicts that a female will do its uttermost best to research both the genetic fitness (e.g., the peacock cock has managed to outwit all his enemies so far, hence he must be strong and smart) and the presence of other favorable properties (e.g., primates' loyalty and care for its offspring) of a male, thereby maximizing the survival of her and her descendants.

7.2. The (Meta)Data Counterpart

If we take a closer look at the MPEG multimedia standards *family of species*, the relevance of specific parts of the MPEG-based encoding standards, *i.e.*, MPEG-2 and MPEG-4, in media and content production as well as in content management is significant. MPEG-2 was the video imagery standard that came into every household TV-set since mid 1980s, whereas its successor MPEG-4 (in different versions) is the current video imagery standard that comes into every iDTV-set and/or (mobile) computer appliance since mid 2000s. The role of the MPEG-based non-coding standards, *i.e.*, MPEG-7 and MPEG-21, on the other hand is more complex, and is harder to assess what impact they have in these domains [18].

The focus of the *multimedia content description interface* specified by MPEG-7, for example, is on providing a comprehensive set of schemes and tools accompanied by the specification of the necessary description language, reference software, conformance guidelines, and extensions [19]. It was one of the main objectives of MPEG-7 to provide a comprehensive and widely applicable set of specifications and tools not limited to any specific domain or content type. As such, they ended up with a huge specification consisting of about 1200 classes. Within current media production, media workflow, media content management, and media archiving systems in the market, *i.e.*, the *picky females*, almost none choose or implemented this MPEG-7 standard, as is was *by nature* far too generic, too verbose, and basically too complicated to use. The Dublin Core (DC) Initiative [20], on the other hand, understood that to make metadata easily understandable and interchangeable, one can already jump quite far by just modeling the 4W's (Who, What, Where, and When) in the right way. As such, DC came up with a very simple and crisp standard of just 15 core classes, which in most cases is enough to be the greatest common divisor between current heterogeneous media metadata sets out there. Eventually, DC took the *descriptive* metadata world by storm in just half a decade. It is needless to say, however, that other standards will prevail, and will take over if one way or another their *genetic fitness* is considered superior. If future media systems see more benefits in another scheme, they will inevitably choose that one over DC [21].

8. Proposition VII: Mutability of the Genome

8.1. Darwin's View

The process that Darwin describes is only sustainable if there are always rivalling genes present within a population. We not only know that his assumption was correct, but by now we can easily point out the origin of this genetic variability, *i.e.*, gene mutations. Mutability is an intrinsic characteristic of genes. To survive, genes have to constantly copy themselves and transfer these copies into new bodies. Genes can copy themselves with amazing accuracy. Nevertheless, the copying process is far from perfect. Approximately one in 10 billion copies contain an error. Such copy errors are called spontaneous mutations. The probability that one particular gene mutates, is small. However, because every human individual consists of some 60 trillion cells, mutations occur regularly [5]. As such, many new genes are launched and tested by nature generation after generation. Our genome contains all the information for the construction of our body in the form of a chemical code, *i.e.*, our 46 chromosomes. In this matter, the environment acts as a source of information for the genome. It continuously brings—as it were—in a brutal, but very effective way information about itself within the genome: to live or let die. By killing generation after generation those individuals that do not have the appropriate genes to survive in a particular environment, the environment automatically describes itself within the genome. As such, all living organisms always reflect both in their construction and operation the true characteristics of the constantly changing natural environment.

8.2. The (Meta)Data Counterpart

It is fair to say that the Internet paradigm of the early nineties redefined Information Technology and information economy as never before. The HyperText Transfer Protocol (HTTP) and the HyperText Markup Language (HTML) proved to be a "deadly" team that took the Information Technology world by storm within a decade. Again, Tim Berners-Lee states that the essential property of the WWW is its universality [1]. The power of a hypertext link is that "anything can link to anything". Web technology, therefore, must not discriminate between the scribbled draft and the polished performance, between commercial and academic information, or among cultures, languages, media, and so on. As such, information varies along many axes. One of these is the difference between information produced primarily for human consumption and that produced mainly for machines. To date, however, that Web has developed most rapidly as a medium of documents for people rather than for data and information that can be processed automatically. The Semantic Web is an evolving development of the WWW in which the meaning (semantics) of information and services on the Web is defined, making it possible for the Web to understand and satisfy the requests of people and machines to use that Web content. It's Tim Berners-Lee's renewed semantic vision of the Web as a universal medium for data, information, and knowledge exchange [22] that changed the nature of information significantly in the last two decades. Now, information is multimedia on the Internet, sensitive to its spatio-temporal roots, live, and dynamic. Furthermore, with the emergence of both citizen-based media and social media, which provide on-line access to different sources and services for commenting and debating on information resources, and use social media to instantaneously (e.g., Twitter) spread new or updated information. This results in large amounts of (possibly) unreliable and repeated information, leaving the user exploring on their own to try to build their own version of an information event from large amounts of potentially related information, or simply to find the truth in the middle of an ocean of rumors and hoaxes. The ultimate goal is to create an environment that facilitates end-users in seeing meaningful connections among individual pieces of multimedia information (stories, photos, graphics, and videos) through underlying knowledge of the descriptions of these individual information items, their relationships, and related background knowledge. This can be solved by semantic metadata models improving metadata interoperability along the entire information chain that leads to individual knowledge and understanding.

Let's connect this insight to our metadata handling of the last half of the millennium again. Throughout the centuries the *mutability of the genome* adhered to the *changing* requirements of the data and information *environment*, as in the early days syntax was enough (e.g., the card index system) and since the dawn of the digital era syntax and easy adjustment was enough (hence the proprietary metadata formats). However, since the advent of the Internet data streams became so big, not only syntax and easy adjustment were "fit" genes, but also easy data exchange became an important differentiator, hence the major breakthrough of standardized XML. Finally, current social information use, coincided with the ever-increasing tsunami of multimedia data, makes the use computing infrastructure inevitable. We need the help of automatic semantic data reasoners that—given one piece of information—can recommend us related linked pieces of data, information, and/or knowledge. This can thus only be done, if we have a metadata standard that on top of syntax, adjustability, and exchangeability also incorporates *semantics*, hence the current hype of triplifying data into RDF—a recent major *mutation* [23].

9. Proposition VIII: Cooperation of Mutations and Natural Selection Leads to Adaptation

9.1. Darwin's View

Coincidence alone can never produce order or intent. However, the cooperation between mutations (*i.e.*, randomness) and natural selection (*i.e.*, necessity) can. Richard Dawkins—perhaps the most perspicacious evolutionary biologist of the current generation—has even demonstrated that a simple computer program combining a random source with a selector brings forth valid English sentences within a very short time span, and after some time even sentences from Shakespeare's Hamlet, *i.e.*, the somewhat funny phrase "Methinks it is like a weasel". Darwin suggested—somewhat reluctantly as a devoted Christian—that nature does not design living organisms with a specific goal in mind [2]. Nature has no purpose. Evolutionary adjustment only occurs by killing everything that is propagating into the faulty direction, and thus by only preserving that small minority that by pure chance stumbles into the right direction. There is no plan, no foresight, no destination, no benevolence: Nature just creates through death and destruction. It is the mere collaboration of those two mutually reinforcing processes that explains why living organisms are generally so well adapted to their natural environment.

9.2. The (Meta)Data Counterpart

In order to *adapt* to this ever increasing crowded world of data and information, as said, we need the help of computers that are able to interpret such vast information streams, thus we have to explain in detail the relation of information to meaning, *i.e.*, "semantics", of the data in the information streams to these machines [1,22]. The only way to do so, is if we can guarantee to have *more and better metadata* by having (a) tools to clean up and reconcile the current metadata; (b) by automatic enrichment of our metadata; (c) by harvesting and exchanging the right metadata; and (d) by consolidating all this metadata for future use. Those four mutually reinforcing processes explain why *living and surviving* (meta)data become so well adapted to this natural evolving information environment.

Metadata clean-up and reconciliation: Before asking the question how to link (meta)data from different sources, we need to develop strategies to check their initial quality and eventually solve issues which might disturb the reconciliation process amongst different resources. We have to acknowledge that there are no established methodologies or tools for metadata quality evaluation, or to put it more bluntly in the words of Diane Hillmann: "There are no metadata police out there" [24]. This is primarily important within social media where free-text tagging is still used a lot. The clue here is to use SKOS folksonomies [25] where only "checked" terms can be used that afterwards can be mapped to known concepts, thereby eliminating, e.g., all kinds of typos. We illustrate within our FreeYourMetadata initiative [26,27] with the help of Google Refine [28] how a quick overview of the metadata quality of a collection can be gained and which type of cleansing and reconciliation actions can be taken (semi)automatically. The cleaning part helps with de-duplication, atomization, blank values, formats and case inconsistencies, and clustering. Afterwards, the reconciliation part maps metadata concepts in a certain (often situation-specific) vocabulary to another (often more commonly used) vocabulary. In case the latter vocabulary forms part of the Semantic Web, this reconciliation annexes the metadata to the Linked Data cloud [29]. Subsequently, machines can now access and interpret these metadata, based on previously acquired knowledge. Reconciliation therefore plays a crucial role in the public availability and dissemination of metadata.

Metadata enrichment: Getting more metadata automatically—an enrichment phase—can be done, e.g., via automatic data analysis, data mining, and/or data understanding, e.g., automatic feature extraction or automatic tag linking via thesauri. One can apply linguistic processing on the plain text contained into some elements of the metadata, the most obvious suspects being author, title, caption, and/or description. The linguistic processing consists of extracting named entities, such as persons, organizations, companies, brands, locations, and events, e.g., via the OpenCalais [30] infrastructure. Once the named entities have been extracted, we could map them to formalize knowledge on the Web available in GeoNames [31] for the locations or in DBPedia [32]/FreeBase [33] for the persons, organizations, and events. Each person entity is therefore mapped to its URI in DBPedia that provides (i) a unique identifier for the resource and (ii) formalized knowledge about this person, such as his biography, career, and genealogy in multiple languages. Therefore, the use of, e.g., the OpenCalais Web service allows us to populate the knowledge base by providing a list of possible instances for all named entities discovered. Furthermore, one could integrate extracted media metadata from shot segmentation tools, scene detection tools, and face recognition tools to be able to also further enrich these extracted named entities [34].

Metadata harvesting and exchange: To further automate and open up publication of information, solutions to harvest and exchange the right data aggregations are needed. We developed such an Open Source framework, *i.e.*, TheDataTank [35], as a framework to open up data in a RESTful and simple way so that it can be easily used in end-user applications, publications or visualizations. TheDataTank is a data harvester, adapter, and aggregator between data owners and the consumers of this data (e.g., app developers). This platform publishes local data on the fly as a simple web API. This makes the original data directly and remotely usable for developers through a universal adapter. This maximizes the potential of your data, while the extra effort of the user is minimized. Each data file is kept at its original location and published through a RESTful API in several popular data formats (e.g., JSON, XML, RDF). This way, the data also remains up-to-date and we can keep track of the usage, as TheDataTank logs all the API calls, which gives you insight in how much, when, where and by whom your data is used. This framework also hooks datasets into the Semantic Web and as such publishes "four star" open data. By using the ontology functionality, the model of a dataset can be described as an ontology as well. Furthermore, we can map its concepts to other ontologies published on the Web, creating interoperability with other datasets. This ontology is published through TheDataTank's REST API as well, so others can use or update it accordingly. A graphical Web application The Semantifier [36] is specifically designed to interact with this ontology API. Together with the REST URI's and the annotations, TheDataTank produces an RDF representation of your data in the popular notations N3, RDF/XML, N-Triple and Talis RDF/JSON.

Metadata provenance: Current information aggregator solutions lack the means to uniquely identify still developing pieces of information, nor do they provide means to do trust estimation on the provenance of all their gathered pieces of information, such that citizens could build a complete and weighted picture of the assembled information for themselves. In order to do so, one needs to solve (a) the issue of getting persistent URI's per developing piece of information; (b) the issue of getting valid, lasting temporal enrichments of versioned pieces of information; and finally (c) the issue of openly publishing provenance information of developing pieces of information on the Web to enable automatic trust estimation. Our proposed architecture is able to do automated trust estimation on developing open pieces of information by extending the Memento paradigm [37] with linked provenance information [38]. To publish these elaborated versions of an information fragment resource and their accompanying provenance information, our platform relies on the Memento datetime content negotiation. We extended this framework to also include HTTP provenance header links for automated discovery of the provenance information. This approach allows us to disseminate the versioned information of the information fragment resources on persistent URIs, depending on the datetime content negotiation to redirect to the appropriate version/memento of the original information fragment resource. Combining datetime content negotiation with the publication of the provenance information—as PREMIS OWL [39], links the provenance information to the datetime dimension of a certain information fragment resource. It even also allows us to store the enrichments of the linked published and preserved information fragment resources, because the temporality of these enrichments is also preserved. Finally, our framework allows discovering the provenance information of the other existing versions of an original information fragment resource bringing provenance information—and thus a possible calculable trust estimation—openly to the Web.

10. Proposition IX: Mankind and Its Behavior are Products of an Evolutionary Process

10.1. Darwin's View

In his *Descent of Man* [3], also from 1871, Darwin openly defended for the first time the idea that man must be related to the current living anthropoid apes, like chimpanzees or orangutans, and that our species and apes have both descended from common ancestors. We now know that there is indeed a deep kinship between all living species on earth. The instructions for the construction of all living organisms are written in the same language. The DNA of both humans and bacteria for example contain the same four nucleotides: adenine, cytosine, guanine, and thymine. Furthermore, Darwin was also convinced that our behavioral characteristics follow the same patterns as the anatomical, morphological, or physiological characteristics of any species. In 1872, he concluded in his *The Expression of the Emotions in Man and Animals* [4] that also the human brain is equipped with behavioral programs through evolution. Darwin understood that emotions (such as, e.g., fright) automatically trigger programs that are carved into our nervous system through evolution. Fright is the typical reaction to a dangerous situation, to which only two possible solutions can be inferred: flight or fight. When frightened your heart automatically beats faster (hence, tissues will receive more energy), your breathing is faster (hence, promoting the uptake of oxygen in your blood and thus building up extra kinetic energy), there is a drastic redistribution of the blood (to only those organs that can optimize your muscular efforts), and one begins to sweat (hence, our ventilation system is turned on in advance). In short, without any willful intervention our body is in the highest state of readiness to maximize our chances for survival, either by fleeing or fighting. Darwin realized that not only the expression of emotions is based on programs in our brain approved by evolution, but that the same applied to the way we reason, the functioning of our memory, the way we perceive the world, *etc.* As such, the whole of human psychology is also a mere reflection of our evolutionary history on this planet. Our evolutionary history is, in a very real sense, *carved* into our mind.

10.2. The (Meta)Data Counterpart

For long, linguists and computer scientists have been trying to construct a comprehensive ontology of the world, enabling automated reasoning tasks on human-structured data. Vital to the success of such an ontology are the number and nature of the relationships between different concepts. In recognition of this importance, the Semantic Web community formed a Linked Open Data (LOD) movement [40], which strives to publish interlinked data in a structured format. Tim Berners-Lee [41] has put forward a five-star scheme to score data quality against. The final, perfecting level can be reached when data is linked against other sources, implying the use of well-defined relationships such as equivalence, inclusion, inheritance, *etc.* As a result, machines are able to both broaden and deepen their understanding of data, since links provide the possibility to look up new data and to relate un-interpreted data to well-understood concepts. From now on machines can come up with newly generated knowledge themselves, as the true information is now already *carved* into the interlinked data clouds on the Internet.

Reconciliation—as stated before—can therefore play a crucial role in the public availability and dissemination of interlinked metadata. The reconciliation part maps metadata concepts in a certain (often situation-specific) vocabulary to another (often more commonly used) vocabulary. In case the latter vocabulary forms part of the Semantic Web, *i.e.*, SKOS thesauri, this reconciliation actually fulfills the fifth star in the Linked Data scheme, as it annexes the metadata to the Linked Data cloud. Subsequently, machines can now access and interpret these metadata, based on previously acquired knowledge.

This Linked Open Data paradigm again opens up new possibilities to generate information in a more specialized way, *i.e.*, (a) domain experts can now model their (often narrow) problem domain very crisply, resulting in extremely specialized ontologies, which could very easily be linked to the LOD-cloud by one single relationship; (b) as it is easy to model and link very specialized information sets, agile solutions pop-up more frequently—e.g., HTML's family of micro-formats—and thus further enrich the *gene pool* of specific problem solutions, often generating extra levels of information; and (c) helper thesauri pop-up, thus helping in the reconciliation phase to automatically end up with better metadata through classification.

All these will help us to elevate the *search as we know it*, as it is apparent that within this information universe, new inherently *carved* semantic search paradigms (e.g., facetted search, tag clouds, multi-dimensional timelines, maps, coverflows, *etc.* through cumulative techniques in autonomous recommendation engines) must prevail to help each one of us improving our knowledge and truth worthiness within each subdomain of our semantically interconnected data and information pools. Albeit, practical implementation might change overtime, the semantic metadata knowledge will continue to evolve, be it in a hereditary way.

11. Discussion

What we learned from Darwin is that finding the "right" information is a significant evolutionary process itself. Over the centuries, we ingeniously coped with ever increasing amounts of data and we will do so in the centuries to come. What I have shown in this article is that the clue to do so lies in the way we capture and treat the metadata thereof and that the use and definition of the metadata evolves in a Darwinian way. More information out there gives birth to more ingenious ways to filter the right information for you. As for now, descriptive metadata is needed to fuel the web crawlers of search engines to build their indexes. However, if these descriptive metadata become intelligent pieces of information themselves (*i.e.*, semantic RDFa), the search engines can better recommend information to the end-user. Adding yet other extra levels of meta-information (*i.e.*, a trail of provenance information) further enriches and better determinates all pieces of information. As such "smart" search engines will be able to convince the end-users to trust certain pieces of information more than others, as this provenance meta-information will give them a calculable proof. At the same time, other pieces of provenance metadata make the data timelessly hereditary in a robust way. This way, search engines themselves also follow Darwinian evolution and it's all about the use of metadata and its evolution thereof.

12. Conclusions

We have demonstrated that Darwin's nine evolutionary propositions show remarkable parallels to both metadata's different modes of information and its significant relation of information to meaning over time. As mankind and its behavior are products of an evolutionary process, the evolutionary process of metadata with its different modes of information is truly on the verge of a new-semantic-era. The overabundance of data on the Internet makes all data in the end deeply intertwined [42]. In an important sense, there are no "subjects" at all or "subjects" too much, *i.e.*, there is only all knowledge or no knowledge, since the cross-connections among the myriad topics of this data world simply cannot be divided up neatly anymore. It is the metadata that can both act as noise-reductors for detailed recommendations to the end-users, as it can be a contributing catalyst to interconnect related information. Over time, metadata thus not only has had different modes of information, but also their relation of information to meaning [43], *i.e.*, "semantics", evolved. It is without any doubt clear that links between things—*i.e.*, metadata—are sometimes as important as, if not more important than, the things—*i.e.* data—themselves. In a sense, metadata expresses helpful *knowledge* about data and though its current semantic interoperability a simple *search for information* becomes in fact truly *querying of metadata knowledge* [44].

Acknowledgments

Furthermore, I would like to thank Jan de Laender (1941–2004) for his valuable insights into Darwin's theories. The research activities that have been described in this paper were funded by Ghent University, Université Libre de Bruxelles, iMinds, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), and the European Union.

References

1. Berners-Lee, T.; Hendler, J.; Lassila, O. The semantic web. *Sci. Am.* **2001**, *284*, 34–43.
2. Darwin, C. *On the Origin of Species*; Murray: London, UK, 1859.
3. Darwin, C. *Descent of Man, and Selection in Relation to Sex*; Murray: London, UK, 1871.
4. Darwin, C. *The Expression of the Emotions in Man and Animals*; Murray: London, UK, 1872.
5. de Laender, J. *Het Verdriet van Darwin. Over de Pijn en de Troost van het Rationalisme*; ACCO: Leuven, Belgium, 2004.
6. Weinberger, D. Metadata and understanding. *KMWorld Magazine*, 29 September 2006, Available online: http://www.kmworld.com/Articles/News/News-Analysis/Metadata-and-understanding-18278.aspx (accessed on 7 December 2012).
7. Raup, D.M. Extinction from a paleontological perspective. *Eur. Rev.* **1993**, *1*, 207–216.
8. ISO/IEC. Space Data and Information Transfer Systems—Open Archival Information System—Reference Model. Available online: http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683 (accessed on 4 December 2012).

9. Mendel, G.J. Versuche über pflanzenhybriden. *Verh. Naturforschenden Ver. Brünn* **1866**, *4*, 3–47.
10. Rosenthal, D. Bit preservation: A solved problem? In *Proceedings of the 5th International Conference on Preservation of Digital Objects*, London, UK, 29–30 September 2008; pp. 1–7.
11. van Valen, L. Molecular evolution as predicted by natural selection. *J. Mol. Evol.* **1974**, *3*, 89–101.
12. Singhal, A. Introducing the knowledge graph: Things, not strings. Available online: http://googleblog.blogspot.be/2012/05/introducing-knowledge-graph-things-not.html (accessed on 4 December 2012).
13. Garud, R.; Jain, S.; Kumaraswamy, A. Institutional entrepreneurship in the sponsorship of common technological standards: The case of SUN microsystems and JAVA. *Acad. Manag. J.* **2002**, *45*, 196–214.
14. The Apache Software Foundation Home Page. Available online: http://www.apache.org (accessed on 4 December 2012).
15. Züger, M.; Poltier, S.; Volkart, A. Economic challenges of standardization. *Internet Econ. V* **2010**, *1*, 31–54.
16. Schema.org Documentation. Avialable online: http://www.schema.org/docs/documents.html (accessed on 4 December 2012).
17. Cusumano, M.; Mylonadis, Y.; Rosenbloom, R. Strategic maneuvering and mass-market dynamics: The triumph of VHS over betamax. *Bus. Hist. Rev.* **1992**, *66*, 51–94.
18. Angelides, M.; Agius, H. *The Handbook of MPEG Applications: Standards in Practice*; Wiley: West Sussex, UK, 2011.
19. ISO/IEC. Information Technology—Multimedia Content Description Interface—Part 1: Systems (MPEG-7). 2002, Available online: http://mpeg.chiariglione.org/tutorials/papers/IEEEMM_mp7overview_withcopyrigth.pdf (accessed on 4 December 2012).
20. Dublin Core Metadata Initiative. DCMI Metadata Terms. 2008, Available online: http://dublincore.org/specifications/ (accessed on 4 December 2012).
21. Weagley, J.; Gelches, E.; Park, J.-R. Interoperability and metadata quality in digital video repositories: A study of Dublin core. *J. Libr. Metadata* **2010**, *10*, 37–57.
22. Shadbolt, N.; Hall, W.; Berners-Lee, T. The semantic web revisited. *IEEE Intell. Syst.* **2006**, *21*, 96–101.
23. Klyne, G.; Carroll, J. Resource Description Framework (RDF): Concepts and abstract syntax. W3C Recommendation. Available online: http://www.w3.org/TR/rdf-concepts (accessed on 4 December 2012).
24. Hillmann, D.; Phipps, J. Application profiles: Exposing and enforcing metadata quality. In *Proceedings of the 7th International Conference on Dublin Core and Metadata Applications*, Singapore, 27–31 August 2007; pp. 53–62.
25. Miles, A.; Bechhofer, S. SKOS simple knowledge organization system reference. W3C Recommendation. Available online: http://www.w3.org/TR/skos-reference/ (accessed on 4 December 2012).
26. van Hooland, S.; Verborgh, R.; de Wilde, M.; Hercher, J.; Mannens, E.; van de Walle, R. Evaluating the success of vocabulary reconciliation for cultural heritage collections. *J. Am. Soc. Inf. Sci. Technol.* **2012**, in press.

27. Free Your Metadata Initiative. Publish and polish your metadata using google refine. 2012, Available online: http://freeyourmetadata.org/ (accessed on 4 December 2012).
28. Google Refine. A power tool for working with messy data. 2012, Available online: http://code.google.com/p/google-refine/ (accessed on 4 December 2012).
29. Cyganiak, R.; Jentzsch, A. Linking open data cloud diagram. **2011**, Available online: http://lod-cloud.net/ (accessed on 4 December 2012).
30. Open Calais Home Page. Available online: http://www.opencalais.com/ (accessed on 4 December 2012).
31. GeoNames Home Page. Available online: http://www.geonames.org/ (accessed on 4 December 2012).
32. DBPedia Home Page. Available online: http://dbpedia.org/ (accessed on 4 December 2012).
33. Freebase Home Page. Available online: http://www.freebase.com/ (accessed on 4 December 2012).
34. de Sutter, R.; Braeckman, K.; Mannens, E.; van de Walle, R. Integrating audiovisual feature extraction tools in media annotation production systems. In *Proceedings of the 13th IASTED International Conference on Internet and Multimedia Systems and Applications*, Honolulu, HI, USA, 17–19 August 2009; pp. 76–81.
35. TheDatatank Open Source Framework. Available online: http://www.thedatatank.com/ (accessed on 4 December 2012).
36. Semantifier Tool. Available online: http://datatank.demo.ibbt.be/The-Semantifier/ (accessed on 4 December 2012).
37. van de Sompel, H.; Sanderson, R.; Nelson, M.; Balakireva, L.; Shankar, H.; Ainsworth, S. Memento: Time travel for the web. **2009**, Available online: http://arxiv.org/abs/0911.1112 (accessed on 4 December 2012).
38. Coppens, S.; Mannens, E.; Vandeursen, D.; Hochstenbach, P.; Janssens, B.; van de Walle, R. Publishing provenance information on the web using the memento datetime content negotiation. In *Proceedings of the 5th WWW Linked Data on the Web Workshop*, Hyderabad, India, March 2011; pp. 6–15.
39. PREMIS—PREservation Metadata: Implementation Strategies. 2012, Available online: http://www.loc.gov/standards/premis/ (accessed on 4 December 2012)
40. Bizer, C.; Heath, T.; Berners-Lee, T. Linked data—The story so far. *Int. J. Semant. Web Inf. Syst.* **2009**, *5*, 1–22.
41. Berners-Lee, T. Linked data—5 star scheme. 2006, Available online: http://www.w3.org/DesignIssues/LinkedData.html (accessed on 4 December 2012).
42. Nelson, T. *Computer Lib/Dream Machines*; Microsoft Press: Redmond, WA, USA, 1987.
43. Logan, R. What is information? Why is it relativistic and what is its relationship to materiality, meaning and organization. *J. Inf.* **2012**, *3*, 68–91.
44. Mannens, E. Interoperability of Semantics in News Production. Ph.D. Thesis, Ghent University, Ghent, Belgium, March 2011.